Bilingual Sentence Alignment Based on Punctuation Marks

Kevin C. Yeh
g904307@oz.nthu.edu.tw
Department of Computer Science
National Tsing Hua University

Abstract

We present a new approach to aligning English and Chinese sentences in parallel corpora based solely on punctuations. Although the length based approach produces high accuracy rates of sentence alignment for clean parallel corpora written in two Western languages such as French-English and German-English, it does not fair as well for parallel corpora that are noisy or written in two distant languages such as Chinese-English. It is possible to use cognates on top of length-based approach to increase alignment accuracy. However, cognates do not exist between two distant languages, therefore limiting the applicability of cognate-based approach. In this paper, we examine the feasibility of using punctuations for high accuracy sentence alignment. We have experimented with an implementation of the proposed method on the parallel corpus of Chinese-English Sinorama Magazine Corpus with satisfactory results. We also demonstrated that the method was applicable to other language pairs such as English-Japanese with minimal additional effort.

1. Introduction

Recently, there are renewed interests in using bilingual corpus for building systems for statistical machine translation (Brown et al. 1988, 1991), including data-driven machine translation (Dolan, Pinkham, and Richardson 2002), computer-assisted revision of translation (Jutras 2000) and cross-language information retrieval (Kwok 2001). It is therefore useful for the bilingual corpus to be aligned at the sentence level with very high precision (Moore 2002; Chuang, You and Chang 2002, Kueng and Su 2002). After that, further analyses such as phrase and word alignment, bilingual terminology extraction can be performed (Melamed 1997).

Much work has been reported in the literature of computational linguistics studying how to align English-French and English-Germany sentences. While the length-based approach (Church and Gale 1991; Brown et al. 1992) to sentence alignment produces surprisingly good results for the language pair of French and English at success rates well over 96% by sentence, it does not fair as well for alignment of English and Chinese sentences. Work on sentence alignment of English and Chinese texts (Wu 1994), indicates that the lengths of English and Chinese texts are not as highly correlated as in French-English task, leading to lower success rate (85-94%) for length based aligners. Simard, Foster, and Isabelle (1992) proposed using cognates on top of length-based approach to improve on accuracy. They use an operational definition of cognates, which include digits, alphanumerical symbols, punctuations and alphabetical words.

Simard, Foster, and Isabelle (1992) pointed out cognates in two close languages such as English and French can be used to measure the likelihood of mutual translation. Those cognates include alphabetic words, numeric expressions, and punctuations that are almost identical and readily recognizable by the computer. However, for distant language pairs such as Chinese and English, there are no orthographic, phonetic or semantic cognates in existence, which are readily recognizable by the computer. Therefore, the inexpensive cognate-based approach is not applicable to the Chinese-English tasks. Since both of the length and cognate-based methods do not present satisfactory alignment results for distant bilingual pairs, we are motivated to find other alternative evidence that two blocks of texts are mutual translation. It turns out that punctuations can be telling evidences, if we do more than hard matching of punctuation and take into consideration of intrinsic sequencing of punctuation in ordered comparison.

2. Punctuation and Sentence Alignment

We will show that punctuations in Chinese and English mark texts with similar semantic properties, therefore, it is very effective to use them to measure the likelihood of mutual translation for a pair of texts.

Punctuation Translation probability

English Punctuation	Chinese Punctuation	Number of counts	Probability
)	）	4	0.9972
,"	」，	6	0.9564
."	」，	3	0.9164
!	！	38	0.8835
,	，	541	0.8099

Punctuation Fertility probability

Punctuation Type	Number of Counts	Probability
0-1	588.0005	0.225027
1-0	286.001	0.109452
1-1	1698.076	0.649852
1-2	2.466198	0.000944
2-1	0.965034	0.000369
2-2	37.19216	0.014233
3-1	0.314022	0.000121

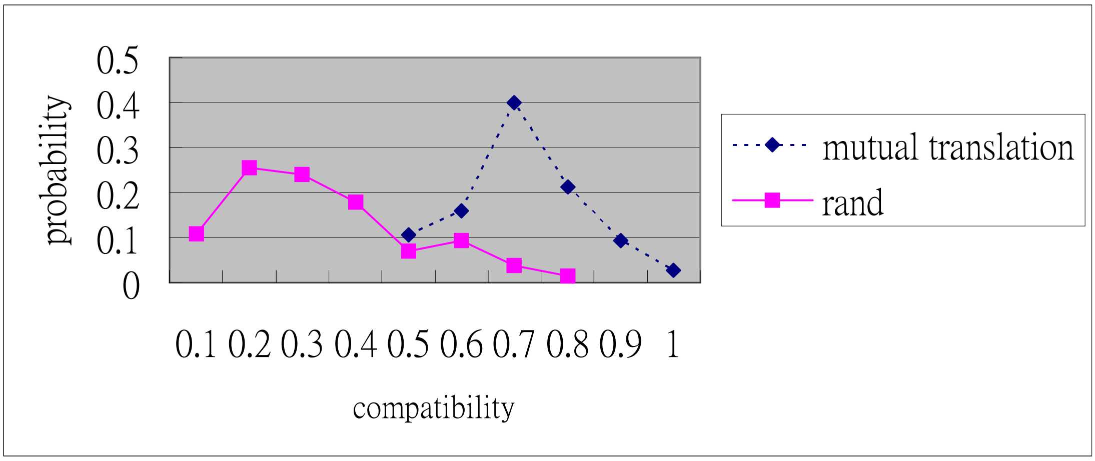
In order to explore the relationship between the punctuations in pairs of Chinese and English sentences that are mutual translations, we selected a small set of manually aligned texts and investigated the characteristics and the associated statistics between the punctuations. Following next a list of the number of counts and the probability relating the Chinese punctuation and the English punctuation was tallied. The information was then used to bootstrap on a larger corpus where an unsupervised EM algorithm and Dynamic programming are used to optimize the punctuation correspondence between a text and its translation counterpart. The EM algorithm converges quickly after the second round of training. Some of the results of the EM training are shown in the above tables.

The probability of the one-to-one match type is about 0.65, which implies that there is a large discrepancy of the punctuation mappings between Chinese and English. The punctuation compatibility is measured by using a relatively larger corpus – around one thousand articles from the Sinorama magazine. Simard, Foster and Isabelle proposed the cognate based approach as a new way of measuring how two pieces of text are mutual translation. The punctuation compatibility as an indicator of mutual translation is defined as

$$\gamma = \frac{c}{\max(n, m)}$$

where γ = punctuation compatibility,
c = number of cognate,
n = number of Chinese punctuations marks
m = number of English punctuation marks

We then took the aligned English-Chinese sentences that have same number of punctuation count (that is the denominator of the equation), take ten counts for example, in order to evaluate how well punctuations work as indicator of mutual translation of English and Chinese sentences. We also took the same English sentences and matched them up with randomly selected Chinese sentences to calculate the compatibility of punctuation marks of unrelated texts.



Results indicated that the average compatibility for pairs of sentences, which are mutual translations, is about 0.67 (with a standard deviation 0.170), while the average compatibility of random pairs of bilingual sentences is 0.34 (with a standard deviation 0.167). The above figure shows compatibility based on punctuation count of ten. The results indicate as the number of punctuations increases the reliability of the compatibility function is more informative. Overall, if the punctuations are soft matched in ordered comparison across the two languages, they indeed provide useful information for effective sentence alignment.

We define a probability of the sequence of punctuations E_i in one language (L1) translating to the sequence C_j of punctuation in another language (L2) as follows:

$$P(E_i, C_j) = \prod_{k=1,m} P(p_k, \pi_k) P(|p_k|, |\pi_k|)$$

where p_k and π_k is one or two punctuations,
$p_1 p_2 \cdots p_m = E_i$, the English punctuations,
$c_1 c_2 \cdots c_m = C_j$, the Chinese punctuations,
$|p_k|$ and $|\pi_k|$ are the number of punctuations in
p_k and π_k respectively,
P(p_k , π_k) = probability of p_k translating into π_k ,
P(*j*, *k*) = probability of *j* punctuations in L1 translating into punctuation in L2.

We observed that in most cases the links of punctuations do not cross each other much like the situation with sentence alignment. Therefore, it is possible to use the dynamic programming procedure to soft match the punctuations across languages, finding the Viterbi path as long as we have the punctuation translation function P(p_k , π_k) and the fertility function P(j, k).

Not like the way Simard et al. (1992) handled cognates, we model the compatibility of punctuations across two languages using Binomial distribution. We model the problem as each punctuation appearing in one language either has a counterpart across translation or not. And for each punctuation, the probability of having a translation counterpart is independent with a fixed value of p.

We differ from Simard approach in the following interesting ways. First, we use the accumulative value of Binomial distribution, while Simard et al. used a likelihood ratio.Second, we go beyond mere hard matching and allow a punctuation mark in one language to match up with a number of compatible punctuations. The compatibility is modeled based on the lexical translation probability proposed by Brown et al. (1991).Finally, we take into consideration of intrinsic sequencing of punctuation in ordered comparison, the flexible and ordered comparison of punctuation is carried out via dynamic programming.

Following Gale and Church (1991), we appeal to Bayes Theorem to estimate the likelihood of aligning two text blocks E and C by calculating P(E, C) P(match).We adopt the same dynamic programming method, but use punctuations to measure the likelihood of mutual translation instead of lengths. For that we define the probability P(E, C) that two text blocks E and C are mutual translation as follows: Given two blocks of text E and C, we first strip off non-punctuations therein to get the punctuations strings E_i and C_j and find out the maximum number of punctuations n. Subsequently, the dynamic programming procedure mentioned before is carried out to find out the value of r, the number of compatible punctuations in ordered comparison of punctuations across languages. Therefore we have:

$$P(E,C) = \sum_{k=1}^{t} P(m_k)P(E_{i,k}, C_{j,k}) = \sum_{k=1}^{t} \mathrm{P}(\mathrm{m}_{\mathrm{k}}) b(r_k, n_k)$$

$$= \sum_{k=1}^{t} \mathrm{P}(\mathrm{m}_{\mathrm{k}}) \binom{n_k}{r_k} p_k^{\,n^k} (1 - p_K)^{r^k - n^k}$$

where n_K = the number of compatible punctuations in ordered comparison,

r_k = the max number of punctuations from English text or Chinese text

p_K = the probability of existence of a compatible punctuation across from one language to the other.

$P(m_K)$ = the match type probability aligning $E_{i,k}$ and $C_{j,k}$

From the data, we have found that about two third of the times, a sentence in one language matches exactly one sentence in the other language (1-1). Other additional possibilities are also considered: 1-0 (including 0-1), and many-1 (including 1-many). Chinese-English parallel corpora are considerably noisier, reflecting from wider possibilities of match types. Here we used the same probabilistic figures as proposed in Chuang and Chang (2002). The following table shows all eight possibilities used in our implementation.

Match type	1-1	1-0, 0-1	1-2	2-1	1-3	1-4	1-5
Probability	0.65	0.000197	0.0526	0.178	0.066	0.0013	0.00013

3.1. First Experiment and Evaluation

In the first experiment, we assessed the performance of punctuation-based sentence alignment, we have randomly selected five bilingual articles from three different bilingual corpora to test out to an implementation of the proposed method. Evaluation of the experiment results were made by native Chinese college students with good knowledge in English. Some experimental results of sentence alignment based on length and punctuation are shown in Appendix (Table A). Shaded parts indicate imprecision in alignment results. We calculated the precision rates by dividing the number of un-shaded sentences (counting both English and Chinese sentences) by total number of sentences proposed. Since we did not exclude aligned pair using a threshold, the recall rate should be the same as the precision rate. The experimental results indicate that when non 1-1 matches next to each other tend to fail the length-based aligner. However, the punctuation-based aligner appears to handle such cases more successfully.

Precision Evaluation using punctuations

Articles	Length-only	Punctuation-only	Improvement
World in a box*	91.5	98.8	7.3
What clones*	86.5	96.6	10.1
New University**	93.0	95.3	2.3
Book I-2 ***	96.5	98.9	2.4
Book II-8 ***	97.1	98.0	0.9

* Scientific American Magazine

** Sinorama Magazine

*** Harry Porter

3.2. Second Experiment and Evaluation

In the second experiment, we evaluated our method testing on a larger range of corpus data. We used all the English and Chinese articles of Scientific American Corpus from January 2003 to December 2003. There are 67 articles, 1523 English sentences, and 1599 Chinese sentences. All the articles include both the English text and their corresponding Chinese counterpart. Here are the results of the experiment:

	Precision	Recall

Excluding partially incorrect and missing errors	95.8%	100.0%
Including partially incorrect and missing errors	93.0%	98.2%

4. Conclusion

We developed a very effective sentence alignment method based on punctuations. The probability of the match between different punctuation marks between the source and the target is calculated based on large bilingual corpora. The punctuation alignment has the property of a binomial distribution. We have experimented with an implementation of the proposed method on a large parallel corpus data. The experiment results show that the punctuation- based approach outperforms the length-based approach with precision rates approaching 93%.

We have explored ways of extending punctuation-based method. First, there is possibility that we may want to interleave the matching of punctuations and regular text segments between punctuations for sub-sentential alignment. We observed that although word alignment links do cross one and other a lot, they general seem not to cross the links between punctuations. Also since the method is quite general, it would be interesting to see if one can adapt the method to handle other language pairs. We have hand-coded a small English-Japanese punctuation mapping table, and convert our alignment program to handle Alignment of Japanese and English texts. It appears that the adapted program works with compatible performance to the original one. Please see example in Appendix (Table B).

A number of interesting future directions present themselves. First, punctuation alignment can be exploited to constrain word alignment and reduce error rates. Second, the punctuation alignment make possible a finer-grained level of bilingual analysis and can provide a strikingly different translation memory and bilingual concordance for more effective example-based machine translation (EBMT), computer assisted translation and language learning (CAT and CALL).

References

Brown, P. F., J. C. Lai and R. L. Mercer (1991), 'Aligning sentences in parallel corpora', in 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA. pp. 169-176.

Chuang, T., G.N. You, J.S. Chang (2002) Adaptive Bilingual Sentence Alignment, Lecture Notes in Artificial Intelligence 2499, 21-30.

Dolan, William B., Jessie Pinkham & Stephen D. Richardson (2002) MSR-MT: The Microsoft Research Machine Translation System, AMTA 2002, 237-239.

Gale, William A. & Kenneth W. Church (1993), A program for aligning sentences in bilingual corpus. In Computational Linguistics, vol. 19, pp. 75-102.

Jutras, J-M 2000. An Automatic Reviser: The TransCheck System, In Proc. of Applied Natural Language Processing, 127-134.

Kueng, T.L. and Keh-Yih Su, 2002. A Robust Cross-Domain Bilingual Sentence Alignment Model, In Proceedings of the 19th International Conference on Computational Linguistics.

Kwok, KL. 2001. NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS. In Proceedings of the Second NTCIR Workshop Meeting, pp. (5) 14-20, National Institute of Informatics, Japan.

Melamed, I. Dan (1997), A portable algorithm for mapping bitext correspondence. In The 35th Conference of the Association for Computational Linguistics (ACL 1997), Madrid, Spain.

Moore, Robert C., 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora., AMTA 2002, 135-144.

Simard, M., G. Foster & P. Isabelle (1992), Using cognates to align sentences in bilingual corpora. In Proceedings of TMI92, Montreal, Canada, pp. 67-81.

Wu, Dekai (1994), Aligning a parallel English-Chinese corpus statistically with lexical criteria. In The Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, USA, pp. 80-87.6. Tainan: National Cheng Kung University.

Appendix

Table A. Experimental result of sentence alignment based on length and punctuation.

Sentence alignment based on length		
Type	English text	Chinese Text
12	Allowing education to be led by the market may also lead to deficiencies in teaching practices.	市場領導教育還可能引發教學上的弊病。台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路，
11	Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams.	「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」
31	"In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams." Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the class expenses fund.	甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。
Sentence alignment based on punctuation		
11	Allowing education to be led by the market may also lead to deficiencies in teaching practices.	市場領導教育還可能引發教學上的弊病。
11	Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams.	台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路，
11	"In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams."	「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」
21	Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the	甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。

	class expenses fund.	

Table B. English-Japanese sentence alignment example.

Sentence alignment based on punctuation		
Type	English text	Japanese Text
12	It turns out that about two-thirds of the names examined were suitable for either women or men. .	その結果、３分の２の名前が男でも女でも通用するものであることがわかった。漢代の女性の名前には実に力強いものも少なくない。 、。。
21	Wang Mang, who usurped the throne in 9 AD, named his daughter Jie ("nimble and quick"). The daughter of the emperor Huan Di (132-167 AD) was named Jian ("solid and resolute") while her mother, the empress Deng, had the even more emphatic name of Mengnu, which means "fierce woman"! **, , (" ") . ( ) (" ") , , , " " !**	王莽の娘の名は「倢」、後漢の桓帝の娘の名は「堅」といい、桓帝の時の皇后の名は、より直接的な「猛女」というものだったのである。 「」、「」、、「」。
11	Says Liu, "These names show that society at that time had not yet come to hold the two sexes to such very different standards." **, " . "**	「この現象は、男性と女性の道徳行為に対する社会の要求が、あまり違わなかったことを示しています」と劉増貴さんは言う。 「、、」。